DOCUMENT RESUME

ED 343 935 TM 018 090

AUTHOR Ferrell, Charlotte M.
TITLE Statistical Significance, Sample Splitting, and Generalizability of Results.
PUB DATE Jan 92
NOTE 28p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Houston, TX, January 31-February 2, 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Equations (Mathematics); Generalization; Mathematical Models; *Predictive Measurement; *Reliability; Sample Size; *Sampling; *Statistical Significance
IDENTIFIERS Criterion Referenced Scoring; Cross Validation; Invariance; *Research Replication; *Sample Splitting

ABSTRACT

Statistical significance is often misinterpreted to mean replicability or generalizability of results, although a statistically significant difference does not equal a reliable difference. Sample splitting procedures may be a more accurate way of estimating research result generalizability. This type of cross-validation involves randomly dividing the original group and using one subgroup to develop the predictor equation. This equation is used to predict a criterion score for each person in the second subgroup, and predicted criterion scores for subjects in the second subgroup are correlated with actual criterion scores. To illustrate the sample splitting invariance procedure, a small data set (n=42 inservice teachers) from a study by L. G. Daniel and K. R. Okeafor (1987) is used. Results demonstrate that statistical significance testing may not always yield an accurate estimate of replicability when compared to analyses run using the sample splitting invariance procedure. There are five tables of analysis results and an appendix listing relevant computer codes for the computations. A 24-item list of references is included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED343935

STATISTICAL SIGNIFICANCE, SAMPLE SPLITTING, AND GENERALIZABILITY OF RESULTS

Charlotte M. Ferrell

University of Southern Mississippi

Hattiesburg, Mississippi

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

CHARLOTTE M. FERRELL

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the annual meeting of the Southwest Educational Research Association, January 30, 1992, Houston, TX.

TM018090

BEST COPY AVAILABLE

ABSTRACT

Contrary to the misnomer that statistical significance implies replicability or generalizability of results, the author illustrates how sample splitting procedures may be a more accurate way of estimating result generalizability. A small data set from an actual study is used to illustrate how statistical significance testing may not always yield an accurate estimate of replicability when compared to analyses run using sample splitting "invariance" procedures. Examples are presented using multiple regression procedures.

Statistical Significance, Sample Splitting, and Generalizability of Results

Researchers often misinterpret a statistically significant result by assuming that it is, by necessity, an important or substantively significant result. This fallacy was noted by Morrison and Henkel (1969) who asserted that

> knowledge of <u>only</u> the level of significance of a sample statistic tells <u>nothing</u> about the magnitude of the relationship or difference being studied, nor does it provide any clues as to its theoretical or other interpretation (p. 32--emphasis in original).

Holmes (1990, p. 72) suggests that this problem may stem from the tendency of many researchers to refer to "statistically significant" results as "significant" results, implying that if a statistically significant difference is found, such a difference is important and noteworthy. More appropriately, Craig, Eison, and Metze (1975, p. 282) assert that ". . . psychology's task is not one of identifying 'significant' differences but 'meaningful' relationships among variables." Gold (1969, p. 44) concluded that ". . .in the controversy over the use of tests of significance, there is no disagreement that the confusion of <u>statistical significance</u> with <u>substantive importance</u> is a sin." (emphasis added)

Selvin's (1957) article criticizing the use of significance tests in survey research generated much controversy and sparked a chain of reactions from noted critics in the field (Morrison & Henkel, 1969). Almost 50 years later, the issues of this debate

have not been resolved. Shaver (1985, p.58) illustrates the problem of misinterpretation of statistical significance in the following amusing dialogue about significance testing:

Chris: [Sitting back again.] Well, that's part of what's so great about my study. Students in my treatment group had a higher mean score on the individually administered test, and the difference was statistically significant.

Jean: Terrific! Was the difference large enough to be important?

Chris: [Looking puzzled.] Well, as I said, it was statistically significant. You know, that means it wasn't likely to be just a chance occurrence. I set the level of significance at 0.05, as my advisor suggested. So a difference that large would occur by chance less than five times in a hundred if the groups weren't really different. An unlikely occurrence like that <u>surely</u> must be important.

Jean: Wait a minute, Chris. Remember the other day when you went into the office to call home? Just as you completed dialing the number, your little boy picked up the phone to call someone. So you were connected and talking to one another without the phone ever

ringing. . . What do you suppose the probability would be of his picking up the phone just as you completed dialing?

Chris: Gee, I couldn't even estimate, but it would have to be minuscule.

Jean: Well, that must have been a truly important occurrence then?

Chris: . . . . . .

Many attempts have been made by critics of statistical methodology to clarify the meaning and interpretation of statistical significance tests (e.g., Carver, 1978; Clark, 1963; Gold, 1969; Morrison & Henkel, 1969; Thompson, 1987, 1988, 1989). Yet researchers still have misconceptions regarding the proper use and interpretation of these tests. Moreover, textbook writers even provide erroneous information regarding the interpretation of statistical significance tests. Consequently, Carver (1978, pp. 378-379) has argued that

> Statistical significance testing has involved more fantasy than fact. The emphasis on statistical significance in educational research represents a corrupt form of the scientific method. <u>Educational research would be better off if it stopped testing its results for statistical significance.</u>
>
> . . . Unfortunately, only a small proportion of the specialists in education who use and interpret statistical significance testing understand what it is

> or how to discriminate between what it is and what it is erroneously interpreted to be. (emphasis added)

Similarly, Clark (1963, pp. 466, 469) suggests that

> The null hypothesis of no difference has been judged to be no longer a sound or fruitful basis for statistical investigation. . . . Significance tests do not provide the information that scientists need, and, furthermore, they are not the most effective method for analyzing and summarizing data.

By definition, the term "statistical significance" simply means that there is a certain probability that the results of a study are due to chance alone (Holmes, 1990, p. 71). A test of statistical significance allows the researcher to make a decision between the "null" and "alternative" hypotheses concerning some characteristic of a population (parameter) based on knowledge gained from a sample (sample statistic) of that population (Morrison & Henkel, 1969). In the social sciences, the null hypothesis is usually "an assertion that variables are unrelated or that groups do not differ with regard to some characteristic" (p. 131).

The decision to accept or reject the null hypothesis is based on the researcher's choice of alpha which is the probability of a Type I error (i.e., the probability of rejecting a true null). The 0.05 level is generally used, but Huberty (1987) points out that no alpha value is sacred. Huberty also

differentiates between alpha values and p values. Researchers often erroneously use the terms synonymously.

For example, an alpha level of 0.05 means the researcher is willing to risk being wrong one out of every 20 times the decision to reject the null hypothesis is made. On the other hand, the p value is the approximate probability that a sample statistic of at least the size observed in the sample would be obtained if the null hypothesis were true (Arney, 1990, p. 202). Generally, if the p value is less than alpha, the null hypothesis is rejected, and if p is greater than alpha the null hypothesis is accepted.

One common criticism of statistical significance testing is that "test statistics often presume random sampling from the population and random assignment to treatment conditions" (Thompson, 1987, p.8). Selvin (1957) has argued that both random selection from a total population and random assignment are necessary for the legitimate application of significance testing. However, other researchers (e.g., McGinnis, 1959; Thompson, 1987) suggest that this premise is not entirely true. Huberty (1987) notes that random sampling does not necessarily insure that the response measures will be independent. McGinnis (1959, p. 413) asserts that "no test of significance requires of itself that all correlated biases be removed, that is, that randomization be effected." In other words, it is acceptable to obtain an

ordinary probability sample from a population without violating the conditions of "the test."

According to Thompson (1987), it has been correctly noted that social science researchers "rarely sample from a total population" and randomly assign subjects to "meaningful experimental groups." However, Thompson argues that significance testing is restricted only to a representative sample of a population which may be obtained by "comparing known sample characteristics with known population characteristics to build some warrant for an assumption of representativeness" (p. 9). In fact, the researcher may wish to consider the set of observations as a "population" rather than a sample. This is an option when a reasonable population cannot be specified if the observations were considered a sample (Gold, 1959).

The only truly legitimate purpose of a significance test is "that of assessing the sampling error of a statistic designed to describe a particular population on the basis of a probability sample" (Morrison & Henkel, 1969, p. 133). According to Huberty (1987, p. 7), ". . . there is nothing wrong with statistical tests themselves," but they should be used as guides and indicators instead of a means of arriving at conclusive answers. Kirk (1982) echoes these sentiments, arguing that

> A test of significance yields the probability of committing an error in rejecting the null hypothesis. It embodies no information concerning the importance or

> usefulness of the result obtained. It is just one bit of information used in making a decision concerning a scientific hypothesis (p. 42).

Therefore, significance testing in all its forms, need not be abandoned (Thompson, 1987).

Although "significance tests [merely] evaluate the calculated probability that a sample result originated in a population in which the null hypothesis is true" (Thompson, 1987, p.11), the published literature generally reflects a bias against statistically nonsignificant results. When results are not statistically significant, authors tend to not submit the studies for publication because editors tend to perceive articles more favorably when statistically significant results are reported (Thompson, 1988). In fact, Greenwald (1975) observed that "very little is expected to be published on problems for which the null hypothesis is, to a reasonable approximation, true."

Schafer (1991) suggests that nonsignificant results simply indicate that the data are not sufficient to estimate a parameter. Consequently, nonsignificant results should be reported and accompanied by a measure of statistical power.

According to Savage (1957, pp. 332-333), "null hypotheses of no difference are usually known to be false before the data are collected." It has been pointed out by Hays (1981, p. 293) that "<u>Virtually any study can be made to show significant results if one uses enough subjects</u>." (emphasis in original)

Statistical Significance and Result Generalizability

Statistical significance is often misinterpreted to mean replicability or generalizability of results. However, a statistically significant difference does not mean "a reliable difference" (Gold, 1969) even though "statistically significant" is frequently interpreted as the probability of obtaining the same results whenever a given experiment is replicated (Carver, 1978). Carver (1978, p. 386) disputes this interpretation, noting that

> nothing in the logic of statistics allows a statistically significant result to be interpreted as directly reflecting the probability that the result can be replicated. It is a fantasy to hold that statistical significance reflects the degree of confidence in the replicability or reliability of results.

Replication is the "ultimate test of scientific findings" (Thompson, 1987, p. 27), and many researchers have suggested that replicability should be used in place of significance testing as part of a "new logic of truth testing." The validity of a proposition in social science research comes from repeated observations, not from random sampling (Gold, 1969). One study can never provide "convincing support for a causal hypothesis or theoretical statement in the social sciences" (Neale & Liebert, 1986, p. 290). As noted by Morrison and Henkel (1969),

"Generalizability involves considerably more than the relation of the sample of research subjects to some population. Generalization involves research subjects, measurement variables, conceptual variables, and test settings" (p. 143). Subsequent studies must replicate the first study with respect to all four categories and must yield similar results to provide confirmatory evidence of the findings and to extend the range of generalizability with respect to the categories of research subjects and measurement variables.

In fact, it has been suggested that significance testing seems to distract from the awareness that replication is vitally important in science (Thompson, 1987). One strategy for replicating a study is to use "hold out" samples to cross-validate results of the analysis. Cross-validation is used for correlational analyses which tend to capitalize on chance (i.e., multiple correlation, canonical correlation analysis, discriminant function analysis, and factor analysis). This technique (Gorsuch, 1974, p. 297; Thompson, 1981) involves randomly splitting samples in half and empirically comparing the statistical estimators extracted from the two data sets. To cross-validate a sample, invariance coefficients, which capitalize on measurement error, are calculated in addition to alpha coefficients, which capitalize on sampling error. These invariance coefficients are similar to test-retest reliability

coefficients, and are produced by the SPSSx CORRELATIONS procedure (Thompson, 1981).

In multiple correlation, the prediction equation is used to predict criterion scores for a new group of individuals. Even when r is statistically significant and a large percentage of the variance is explained by the predictor variables, the researcher cannot be certain that the regression equation will be accurate with a different group of people (Huck, Cormier, & Bounds, 1974). In fact, the initial equation could be the result of chance (Gay, 1987). Hence, it is recommended that the prediction equation be validated with at least one other group and that variables not related to the criterion measure be removed.

> The technique of cross-validation involves four steps: (1) The original group of people (for whom both predictor and criterion scores are available) is randomly divided into two subgroups. (2) Just one of the subgroups is used to develop the prediction equation. (3) This equation is used to predict a criterion score for each person in the second subgroup (i.e., the subgroup that was not used to develop the prediction equation). (4) The predicted criterion scores for people in the second subgroup are correlated with their actual criterion scores (Huck et al, 1974, p. 159).

If the correlation is statistically significant and if the people

in future studies are similar to those in the cross-validation, then the prediction equation may be used for new groups of people with some degree of accuracy.

It has been noted by Thompson (1981) that the computation of invariance coefficients in addition to alpha coefficients may allow the researcher to generalize results more accurately when correlational methods are utilized. Thompson (1987) warns, however, that cross-validation results warrant cautious interpretation, noting that "all procedures are 'liberal' estimates of invariance when one data set is split into subgroups, because the two subgroups came from one sample and the subgroups and their parameter estimates are therefore interdependent" (p. 29).

Methodology

The data set used in this paper was taken from a study by Daniel and Okeafor (1987). Since the purpose of the present investigation is to illustrate various facets of statistical significance tests, no substantive interpretation of the data is presented.

Analyses were run using nine predictor variables. The variable names and labels are shown in Table 3. The dependent variable is MYPROF (my professionalism) which is one of the subscales of the Logic of Confidence Scale (Daniel & Okeafor, 1987). This scale is used to determine the level of professional recognition teachers feel they personally deserve. In order to

illustrate more effectively the significance issues, only two cohorts of the original three in the data set (Daniel & Okeafor, 1987) were included. The preservice teachers were omitted to obtain a smaller sample (n = 42) in order to more effectively illustrate the invariance procedure explained in this paper. As previously mentioned, these data are being used only for illustrative purposes. If this had been an actual substantive study, it is likely that the methods selected may have been different. No hypotheses about the relationship between professionalism and any of the predictor variables were tested.

The first regression analysis was run using the full inservice teacher sample (n = 42). These data are shown in Table 1. Next, the sample was split based on a contrived "invariance" variable. This was done by randomly assigning cases to two invariance groups. The regression was run again with only invariance group one people and separately with only invariance group two people for a total of three analyses.

Computing Z-Scores

After the two separate analyses for the invariance groups were run, z-scores were computed for all the predictor variables across the two subsets. According to Sprinthall (1990, p. 70), z-scores are ". . . helpful in the interpretation of raw score performances since they take into account both the mean of the distribution and the amount of variability." Therefore, z-scores were computed separately for each of the nine predictor variables

across each of the subgroups using the means and standard deviations, respectively, for each group. This resulted in a new set of z variables for all of the raw score variables. The z-scores were then weighted with Beta weights to compute Y-hat (predicted) values since Beta weights are appropriate when z-scores are used as opposed to a and b weights when raw scores are used in a prediction equation.

Computing Y-hat Scores

Four sets of synthetic scores were computed using these z-scores. The first set of Y-hat values (Y-hat $_{1,1}$) was computed for the independent variables from the first invariance group, using that group's z-scores and Beta weights. The second set of scores (Y-hat $_{1,2}$) was computed using z-scores from the first group and the Beta weights from the second group. Invariance scores for the third group (Y-hat $_{2,1}$) were computed using z-scores from the second invariance group and Beta weights from the first invariance group. Finally, the fourth set of synthetic scores (Y-hat $_{2,2}$) was computed with z-scores and Beta weights from the second invariance group. The SPSSx commands for computing the z-scores and Y-hat values are shown in Appendix A. The predicted invariance scores are obtained by using the opposite set of weights. If the groups are truly invariant, the scores for the opposite group can be predicted using the opposite group's weights.

Computing Invariance Coefficients

Next, Pearson correlation coefficients were computed for the synthetic variables and for the actual dependent variable. These correlations are the invariance coefficients. If the groups are sample invariant, the different sets of synthetic scores will be highly correlated. For example, Y-hat $_{1,1}$ should be highly correlated with Y-hat $_{1,2}$ and Y-hat $_{2,1}$ should be highly correlated with Y-hat $_{2,2}$. Ideally, the correlations across all samples should be high (at least 0.80). However, if this criterion is not met, the sets of weights across one sample should at least be highly correlated. Usually, within the individual sample, the correlations will be higher than they would be across the two samples. As previously mentioned, Thompson (1981) likens these invariance coefficients to test-retest coefficients.

Results

The initial regression results using the entire sample of 42 people are shown in Table 2. Regression results using the first invariance group and the second invariance group are shown in Table 3 and Table 4, respectively. Using only the full inservice sample in the original analysis, results were statistically significant ($p = .0407$) and there was a fairly noteworthy effect ($R^2 = .39208$). However, when the regression analysis was run separately for each of the individual invariance groups, results were not statistically significant, and the Pearson correlations (i.e., the invariance coefficients) shown in Table 5 between the

cross-weighted Y-hat values indicated that the results were not sample invariant. The failure of this sample to produce invariant results may be due in part to the small sample sizes of the two subgroups (n's of 22 and 20).

However, if the original sample had been much smaller, it is possible that the initial results using the entire sample may not have been statistically significant. Had a researcher consulted these original results without benefit of the invariance analysis, the results may have been generalizable simply because they were statistically significant. The invariance results show that in this case the original result was statistically significant merely due either to an artifact of the sample size or possibly to chance.

Ideally, if there is a good regression equation, there will be a high correlation (R^2) between the independent variable composite and the dependent variable. To the extent that this is the case, the results can be used to effectively predict dependent variable scores. When using its own Beta weights to predict Y-hat values, the independent variables for subgroup one yielded fairly accurate predictions of the dependent variable (invariance coefficient = .7638). This was also true for group two when using its own Beta weights (invariance coefficient = .8140). However, when opposite weights were used, none of the correlations were noteworthy (i.e., all expressed less than a five percent correlation).

Invariance is a way of redefining result generalizability. If the results are about the same across the two invariance samples, then there is evidence that the original results are somewhat generalizable to the population of interest. Results of the invariance analyses using the present sample indicate that it would be a fallacy to generalize the original regression results to the population. Hence, as previously noted, it could be argued that the original significance capitalized on chance. Interestingly, even though the invariance analysis showed that the results were not generalizable, the effect size became larger the more the sample was split. In essence, these analyses show how valuable invariance coefficients can be and, conversely, how dangerous it can be to interpret statistical significance as implying that a given result will necessarily be generalizable.

References

Arney, W. R. (1990). Understanding statistics in the social sciences. New York: W. H. Freeman and Company.

Carver, R. P. (1978). The case against statistical significance testing. Harvard Educational Review, 48(3), 378-399.

Clark, C. A. (1963). Hypothesis testing in relation to statistical methodology. Review of Educational Research, 33, 455-473.

Craig, J. R., Eison, C. L., & Metze, L. P. (1976). Significance tests and their interpretations: An example utilizing published research and omega-squared. Bulletin of the Psychonomic Society, 7(3), 280-281.

Daniel, L. G., & Okeafor, K. R. (1987, November). Teaching experience and confidence in teachers. Paper presented at the annual meeting of the Mid-South Educational Research Association, Mobile, AL.

Gay, L. R. (1987). Educational research: Competencies for analysis and application (3rd ed.). Columbus, OH: Merrill.

Gold, D. (1969). Statistical tests and substantive significance. The American Sociologist, 4, 42-46.

Gorsuch, R. L. (1974). Factor Analysis. Philadelphia: Saunders, 1974.

Greenwald, A. G. (1975). Consequences of the prejudice against the null hypothesis. Psychological Bulletin, 82, 1-20.

Holmes, C. B. (1990). The honest truth about lying with statistics. Springfield, IL: C. C. Thomas.

Huberty, C. J. (1987). On statistical testing. Educational Researcher, 16(4), 4-9.

Huck, S. W., Cormier, W. H., & Bounds, W. G. (1974). Reading Statistics and Research. New York: Harper & Row.

Kirk, R. E. (1982). Experimental design: Procedures for the behavioral sciences (2nd ed.). Pacific Grove, CA: Brooks/Cole.

McGinnis, R. (1958). Randomization and inference in sociological research. American Sociological Review, 22, 519-527.

Morrison, D. E., & Henkel, R. W. (1969). Significance tests reconsidered. American Sociologist, 4, 131-140.

Savage, R. J. (1957). Nonparametric significance. Journal of the American Statistical Association, 52, 331-344.

Schafer, W. D. (1991). Power analysis in interpreting statistical nonsignificance. Measurement and Evaluation in Counseling and Development, 23, 2-3.

Selvin, H. C. (1957). A critique of tests of significance in survey research. American Sociological Review, 22, 519-527.

Shaver, J. P. (1985). Chance and nonsense. Phi Delta Kappan, 67(1), 57-60.

Sprinthall, R. C. (1990). Basic statistical analysis (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Thompson, B. (1981). Utility of invariance coefficients. Perceptual and Motor Skills, 52, 708-710.

Thompson, B. (1987). The use (and misuse) of statistical significance testing: Some recommendations for improved editorial policy and practice. Paper presented at the annual meeting of the American Educational Research Association, New Orleans. (ERIC Document Reproduction Service No. 295 957)

Thompson, B. (1988). A note about significance testing. Measurement and Evaluation in Counseling and Development, 20(4), 146-148.

Thompson, B. (1989). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. Measurement and Evaluation in Counseling and Development, 22, 2-6.

Table 1

Listing of Data

ID1	INVAR	LEVEL	GROUP	SEX	ORIENT	YRSEXP1	YRSCOMP	STATUS	DEGREE	WORKAUT	MYPROF
1	1	1	1	2	2	3	2	2	1	39.00	57.00
2	2	1	1	2	1	5	10	2	2	61.00	61.00
3	1	1	1	2	1	6	6	2	1	40.00	50.00
4	2	1	1	2	2	5	6	2	1	47.00	53.00
5	1	1	1	2	1	11	11	1	2	52.00	.
6	2	1	1	2	1	4	15	2	1	59.00	59.00
7	1	2	1	2	2	3	3	2	1	55.00	65.00
8	2	1	1	2	.	16	16	1	2	42.00	.
9	1	1	1	2	1	7	7	1	1	45.00	56.00
10	2	1	1	1	1	33	18	1	4	34.00	40.00
11	1	1	1	2	1	13	15	2	2	69.00	66.00
12	2	1	1	2	1	25	2	1	3	.	61.00
13	1	1	1	2	1	9	4	1	2	48.00	62.00
14	2	1	1	2	1	18	2	1	3	54.00	58.00
15	1	1	1	2	1	6	6	1	2	41.00	63.00
16	2	1	1	2	2	15	11	1	2	41.00	57.00
17	1	1	1	2	1	.	6	1	1	62.00	64.00
18	2	2	1	2	2	11	10	1	1	.	63.00
19	1	1	1	2	1	7	19	2	1	41.00	58.00
20	2	2	1	1	1	29	16	1	2	43.00	64.00
21	1	3	1	2	1	37	38	1	2	60.00	64.00
22	2	3	1	1	1	22	31	1	3	47.00	63.00
23	1	3	1	2	.	15	9	1	2	36.00	48.00
24	2	2	1	1	2	11	11	1	1	59.00	61.00
25	1	1	1	2	2	9	9	1	1	35.00	46.00
26	2	1	1	2	2	5	5	1	1	44.00	50.00
27	1	1	1	2	2	16	16	1	1	45.00	60.00
28	2	2	1	2	1	6	15	1	1	42.00	65.00
29	1	3	1	1	2	25	2	1	2	59.00	62.00
30	2	1	1	2	1	16	7	1	1	57.00	.
31	1	1	1	2	2	5	6	1	1	58.00	57.00
32	2	1	1	2	1	5	6	1	2	55.00	58.00
33	1	3	1	2	1	3	3	2	1	40.00	38.00
34	2	3	1	2	.	20	23	1	3	58.00	62.00
35	1	3	1	2	1	3	3	2	1	47.00	52.00
36	2	1	1	2	2	30	.	1	1	44.00	60.00
37	1	1	1	2	1	17	17	1	3	48.00	65.00
38	2	2	1	2	2	22	22	1	2	59.00	63.00
39	1	2	1	2	1	24	24	1	2	70.00	.
40	2	1	2	2	2	1	2	2	2	53.00	56.00
41	1	1	2	2	2	1	1	2	1	59.00	64.00
42	2	1	2	2	2	1	1	2	1	58.00	53.00
43	1	1	2	2	2	1	1	2	1	38.00	55.00
44	2	1	2	2	2	1	1	2	1	47.00	56.00
45	1	1	2	2	2	1	1	2	2	42.00	51.00
46	2	1	2	2	1	1	1	2	1	51.00	49.00
47	1	4	2	2	2	1	1	2	1	63.00	62.00
48	2	1	2	2	2	1	1	2	1	49.00	54.00
49	1	1	2	2	2	1	1	2	1	41.00	62.00
50	2	1	2	2	2	1	1	2	1	37.00	54.00
51	1	1	2	2	1	1	1	1	1	46.00	63.00
52	2	1	1	2	2	10	11	1	3	54.00	.

Table 2

Regression Using Full Inservice Teacher Sample

Multiple R	.62616
R Square	.39208
Adjusted R Square	.22110
Standard Error	5.84847

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	9	705.92878	78.43653
Residual	32	1094.54741	34.20461

F = 2.29316 Signif F = .0407

Table 3

Regression Using First Invariance Group

	Mean	Std Dev	Variance	Label
LEVEL	1.545	.963	.926	LEVEL OF SCHOOL WHERE TEACHING
STATUS	1.545	.510	.260	EMPLOYMENT STATUS
GROUP	1.273	.456	.208	GROUP BY LEVEL OF EXPERIENCE
ORIENT	1.500	.512	.262	DOES SCHOOL HAVE ORIENTATION PROGRAM FOR
SEX	1.955	.213	.045	
DEGREE	1.364	.581	.338	HIGHEST COMPLETED DEGREE
YRSEXP1	7.955	9.037	81.665	TOTAL YEARS OF TEACHING EXPERIENCE
WORKAUT	48.136	9.473	89.742	WORK AUTONOMY SUBSCALE SCORE
YRSCOMP	7.364	8.937	79.861	YEARS SINCE COMPLETING TEACHER EDUCATION
MYPROF	58.091	7.104	50.468	PROFESSIONALISM SUBSCALE SCORE--I FORM

Multiple R	.76379
R Square	.58337
Adjusted R Square	.27091
Standard Error	6.06594

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	9	618.27122	68.69680
Residual	12	441.54697	36.79558

F = 1.86699 Signif F = .1552

Variables in the Equation

Variable	B	SE B	Beta	T	Sig T
YRSCOMP	.205084	.541606	.257985	.379	.7116
LEVEL	-2.579802	1.805734	-.349527	-1.429	.1786
SEX	-4.357322	13.799852	-.130768	-.316	.7576
ORIENT	.179525	3.376815	.012933	.053	.9585
WORKAUT	.495364	.167138	.660568	2.964	.0118
STATUS	-2.113591	3.670838	-.151630	-.576	.5754
GROUP	3.858137	3.692381	.247563	1.045	.3167
DEGREE	2.027552	3.225489	.165847	.629	.5414
YRSEXP1	-.132579	.668889	-.168649	-.198	.8462
(Constant)	41.615840	30.051024		1.385	.1913

Table 4

Regression Using Second Invariance Group

	Mean	Std Dev	Variance	Label
LEVEL	1.300	.571	.326	LEVEL OF SCHOOL WHERE TEACHING
STATUS	1.450	.510	.261	EMPLOYMENT STATUS
GROUP	1.300	.470	.221	GROUP BY LEVEL OF EXPERIENCE
ORIENT	1.550	.510	.261	DOES SCHOOL HAVE ORIENTATION PROGRAM FOR
DEGREE	1.750	.910	.829	HIGHEST COMPLETED DEGREE
YRSEXP1	9.800	10.014	100.274	TOTAL YEARS OF TEACHING EXPERIENCE
SEX	1.800	.410	.168	
WORKAUT	49.700	7.868	61.905	WORK AUTONOMY SUBSCALE SCORE
YRSCOMP	9.300	8.329	69.379	YEARS SINCE COMPLETING TEACHER EDUCATION
MYPROF	56.900	6.181	38.200	PROFESSIONALISM SUBSCALE SCORE--I FORM

Multiple R	.81399
R Square	.66258
Adjusted R Square	.35890
Standard Error	4.94875

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	9	480.89875	53.43319
Residual	10	244.90125	24.49012

F = 2.18183 Signif F = .1201

---------- Variables in the Equation ----------

Variable	B	SE B	Beta	T	Sig T
YRSCOMP	-.264987	.380624	-.357114	-.696	.5022
WORKAUT	.196943	.170993	.250710	1.152	.2762
ORIENT	.627152	2.614168	.051792	.240	.8153
DEGREE	1.061700	2.330855	.156399	.455	.6585
SEX	6.999425	4.769517	.464760	1.468	.1730
STATUS	.993783	4.835420	.082070	.206	.8413
GROUP	-5.490698	5.293891	-.417680	-1.037	.3241
YRSEXP1	-.055950	.283964	-.090648	-.197	.8478
LEVEL	11.457427	5.050621	1.058948	2.269	.0467
(Constant)	25.497887	17.196581		1.483	.1690

Table 5

Computing Invariance Coefficients

	YHAT11	YHAT12	YHAT21	YHAT22	MYPROF
YHAT11	1.0000	-.0950	.	.	<u>.7638</u>
	(0)	(24)	(0)	(0)	(22)
	P= .	P= .329	P= .	P= .	P= .000
YHAT12	-.0950*	1.0000	.	.	-.1181
	(24)	(0)	(0)	(0)	(22)
	P= .329	P= .	P= .	P= .	P= .300
YHAT21	.	.	1.0000	.2229	.1841
	(0)	(0)	(0)	(21)	(20)
	P= .	P= .	P= .	P= .166	P= .219
YHAT22	.	.	.2229*	1.0000	<u>.8140</u>
	(0)	(0)	(21)	(0)	(20)
	P= .	P= .	P= .166	P= .	P= .000
MYPROF	.7638	-.1181	.1841	.8140	1.0000
	(22)	(22)	(20)	(20)	(0)
	P= .000	P= .300	P= .219	P= .000	P= .

Note: Underlined values are the correlations between the dependent variable and the composite of the independent variable set using its own weights. The asterisks indicate correlations between the two sets of Y-hat scores for a given sample (weighted by original weights and cross-sample weights). These latter correlations are the invariance coefficients.

Appendix A

SPSSx Codes for Computing z-Scores and Y-Hat Values

```
113.500  114  0 SUBTITLE 'COMPUTING ZSCORES AND YHAT VALUES'
114.000  115  0 COMPUTE ZLEVEL1=(LEVEL-1.545)/.963
115.000  116  0 COMPUTE ZSTATUS1=(STATUS-1.545)/.51
116.000  117  0 COMPUTE ZGROUP1=(GROUP-1.273)/.456
117.000  118  0 COMPUTE ZORIENT1=(ORIENT-1.5)/.512
118.000  119  0 COMPUTE ZSEX1=(SEX-1.955)/.213
119.000  120  0 COMPUTE ZDEGREE1=(DEGREE-1.364)/.581
120.000  121  0 COMPUTE ZYRSEXP1=(YRSEXP1-7.955)/9.037
121.000  122  0 COMPUTE ZWORKAU1=(WORKAUT-48.136)/9.473
122.000  123  0 COMPUTE ZYRSCOM1=(YRSCOMP-7.364)/8.937
123.000  124  0 COMPUTE ZMYPROF1=(MYPROF-58.091)/7.104
124.000  125  0 COMPUTE ZLEVEL2=(LEVEL-1.3)/.571
125.000  126  0 COMPUTE ZSTATUS2=(STATUS-1.45)/.51
126.000  127  0 COMPUTE ZGROUP2=(GROUP-1.3)/.47
127.000  128  0 COMPUTE ZORIENT2=(ORIENT-1.55)/.51
128.000  129  0 COMPUTE ZDEGREE2=(DEGREE-1.75)/.91
129.000  130  0 COMPUTE ZYRSEXP2=(YRSEXP1-9.8)/10.014
130.000  131  0 COMPUTE ZSEX2=(SEX-1.8)/.41
131.000  132  0 COMPUTE ZWORKAU2=(WORKAUT-49.7)/7.868
132.000  133  0 COMPUTE ZYRSCOM2=(YRSCOMP-9.3)/8.329
133.000  134  0 COMPUTE ZMYPROF2=(MYPROF-56.9)/6.181
134.000  135  0 IF (INVAR EQ 1) YHAT11=(ZYRSCOM1 * .258) + (ZLEVEL1 * -.350) +
135.000  136  0   (ZSEX1 * -.131) + (ZORIENT1 * .013) + (ZWORKAU1 * .661) +
136.000  137  0   (ZSTATUS1 * -.152) + (ZGROUP1 * .248) + (ZDEGREE1 * .166) +
137.000  138  0   (ZYRSEXP1 * -.169)
138.000  139  0 IF (INVAR EQ 1) YHAT12=(ZYRSCOM1 * -.357) + (ZWORKAU1 * .251) +
139.000  140  0   (ZORIENT1 * .052) + (ZDEGREE1 * .156) + (ZSEX1 * .465) +
140.000  141  0   (ZSTATUS1 * .082) + (ZGROUP1 * -.418) + (ZYRSEXP1 * -.091) +
141.000  142  0   (ZLEVEL1 * 1.059)
142.000  143  0 IF (INVAR EQ 2) YHAT21=(ZYRSCOM2 * .258) + (ZLEVEL2 * -.350) +
143.000  144  0   (ZSEX2 * -.131) + (ZORIENT2 * .013) + (ZWORKAU2 * .661) +
144.000  145  0   (ZSTATUS2 * -.152) + (ZGROUP2 * .248) + (ZDEGREE2 * .166) +
145.000  146  0   (ZYRSEXP2 * -.169)
146.000  147  0 IF (INVAR EQ 2) YHAT22=(ZYRSCOM2 * -.357) + (ZWORKAU2 * .251) +
147.000  148  0   (ZORIENT2 * .052) + (ZDEGREE2 * .156) + (ZSEX2 * .465) +
148.000  149  0   (ZSTATUS2 * .082) + (ZGROUP2 * -.418) + (ZYRSEXP2 * -.091) +
149.000  150  0   (ZLEVEL2 * 1.059
```